09/629,831

1. (Currently Amended) A method of automatically creating a dictionary for clustering text documents comprising:

inputting a maximum dictionary size;

determining a frequency of each word in each of said documents;

creating a dictionary of most frequently occurring words in said documents as limited by said maximum dictionary size, such that said dictionary contains less than all words in said documents;

determining a frequency of phrases in each of said documents that contain only words in said dictionary;

adding most frequently occurring phrases to said dictionary; and

outputting said most frequently occurring words and said most frequently occurring phrases as said dictionary, wherein said dictionary size limits the number of words and phrases maintained in said dictionary.

2. (Previously Presented) The method in claim 1, wherein said determining a frequency of each word comprises:

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

removing duplicate words from said documents;

adding remaining words to said dictionary as limited by said maximum dictionary size;

determining said frequency of each word remaining in said dictionary; and

removing words below a frequency level from said dictionary.

3. (Original) The method in claim 2, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

4. (Previously Presented) The method in claim 1, wherein said determining a frequency of

phrases comprises:

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

adding said phrases in each of said documents that contain only words in said dictionary to said dictionary;

determining said frequency of said phrases remaining in said dictionary; and

removing phrases below a frequency level from said dictionary.

5. (Original) The method in claim 4, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

6. (Currently Amended) A method of automatically creating a dictionary for clustering text documents comprising:

inputting a maximum dictionary size;

performing a first pass for each of said documents comprising:

determining a frequency of each word in each of said documents; and

creating a dictionary of most frequently occurring words in said documents as limited by said maximum dictionary size, such that said dictionary contains less than all words in said documents;

performing a second pass for each of said documents comprising:

determining a frequency of phrases in each of said documents that contain only words in said dictionary; and

adding most frequently occurring phrases to said dictionary; and

outputting said most frequently occurring words and said most frequently occurring phrases as said dictionary, wherein said dictionary size limits the number of words and phrases maintained in said dictionary.

7. (Previously Presented) The method in claim 6, wherein said determining a frequency of each word comprises:

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

removing duplicate words from said documents;

adding remaining words to said dictionary as limited by said maximum dictionary size;

determining said frequency of each word remaining in said dictionary; and

removing words below a frequency level from said dictionary.

8. (Original) The method in claim 7, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

9. (Previously Presented) The method in claim 6, wherein said determining a frequency of phrases comprises:

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

adding said phrases in each of said documents that contain only words in said dictionary to said dictionary;

determining said frequency of said phrases remaining in said dictionary; and

removing phrases below a frequency level from said dictionary.

10. (Original) The method in claim 9, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

11. (Currently Amended) A program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform a method of

automatically creating a dictionary for clustering text documents, said method comprising:

inputting a maximum dictionary size;

determining a frequency of each word in each of said documents;

creating a dictionary of most frequently occurring words in said documents as limited by said maximum dictionary size, such that said dictionary contains less than all words in said documents;

determining a frequency of phrases in each of said documents that contain only words in said dictionary;

adding most frequently occurring phrases to said dictionary; and

outputting said most frequently occurring words and said most frequently occurring phrases as said dictionary, wherein said dictionary size limits the number of words and phrases maintained in said dictionary.

12. (Previously Presented) A program storage device as in claim 11, wherein said determining a frequency of each word comprises:

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

removing duplicate words from said documents;

adding remaining words to said dictionary;

determining said frequency of each word remaining in said dictionary; and

removing words below a frequency level from said dictionary.

13. (Original) A program storage device as in claim 12, further comprising inputting one or more of said stop words, said synonyms, and said frequency level.

14. (Previously Presented) A program storage device as in claim 11, wherein said determining a frequency of phrases comprises:

removing punctuation and case from said documents;

removing stop words from said document;

replacing words in said documents with synonyms;

adding said phrases in each of said documents that contain only words in said dictionary to said dictionary;

determining said frequency of said phrases remaining in said dictionary; and

removing phrases below a frequency level from said dictionary.

15. (Original) A program storage device as in claim 14, further comprising inputting said stop words.

16. (Original) A program storage device as in claim 14, further comprising inputting said synonyms.

17. (Original) A program storage device as in claim 14, further comprising inputting said frequency level.